AD NUMBER
ADB130485
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution authorized to DoD only; Specific Authority; OCT 1988. Other requests shall be referred to Marine Corps., Research Development and Acquisition Command, Washington, DC.**
AUTHORITY
ONR ltr 19 May 1989

CRM 88-171 / October 1988

RESEARCH MEMORANDUM

TWO CONSEQUENCES OF IMPROVING A TEST BATTERY

by D. R. Divgi

A Division of CNA *Hudson Institute*

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

Work conducted under contract N00014-87-C-0001.

CNA

CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

24 October 1988

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 88-171

Encl: (1) CNA Research Memorandum 88-171, "Two Consequences of Improving a Test Battery," by D. R. Divgi, October 1988

1. Enclosure (1) is forwarded as a matter of possible interest.

2. A computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) is being developed for joint-service use. It is likely that many ASVAB subtests will be more reliable in the CAT version than in the present paper-pencil format. Improvement in reliability can increase mean differences between population subgroups and also increase variances of composite scores. This Research Memorandum illustrates these effects through computer simulation and finds their magnitudes to be small.

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 88-171

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN MANPOWER (2 copies)
A2A	CNR
A6	HQMC MPR
	Attn: M
	Attn: MP
	Attn: MR
	Attn: MA (2 copies)
	Attn: MPP-39
A6	HQMC RA
A6	HQMC AVN
A6	CG MCRDAC, Washington
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL (2 copies)
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FKQ6D	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Director, Testing Systems (Code 63)
	Attn: Technical Library
	Attn: Director, Personnel Systems (Code 62)
	Attn: CAT/ASVAB PMO
	Attn: Manpower Systems (Code 61)
FT1	CNET
V12	CG MCRDAC, Quantico
	Attn: Director, Development Center Plans Division (Code D08) (2 copies)
	Attn: Commanding General
V12	CGMCCDC
	Attn: Training and Education Center

OPNAV

OP-01
OP-11
OP-13
OP-15

OTHER

Joint Service CAT-ASVAB Working Group (16 copies)
Defense Advisory Committee on Military Personnel Testing (8 copies)

CRM 88-171 / October 1988

TWO CONSEQUENCES OF IMPROVING A TEST BATTERY

D. R. Divgi

Marine Corps Operations Analysis Group

A Division of CNA *Hudson Institute*

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Replacement of a paper-pencil test battery with a computerized. adaptive version is likely to increase reliabilities of the subtests. This leads to an increase in the variances of composite scores, and to lower mean scores for subgroups whose average scores are already below those of the general population. These results are illustrated with a computer simulation.

EXECUTIVE SUMMARY

INTRODUCTION

The Department of Defense may implement a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery in the near future. If replacement of a paper-pencil (PP) test battery with CAT increases reliabilities of the subtests, the battery is improved. However, such improvement may create some practical problems. One problem concerns the use of composite scores, and the other is the potential adverse impact on minorities whose mean scores are substantially below those of the general population. The purpose of this paper is to illustrate these problems with a computer simulation.

A composite score is a sum of subtest scores, usually after converting subtest raw scores to standard scores. If subtests become more reliable, their intercorrelations increase. This leads to larger variances of composite scores in CATs compared to PP versions, in spite of equating of CAT and PP scores at the subtest level. An appreciable differences between the two variances may necessitate a second equating at the composite level which, being a departure from current practice, would require software changes at agencies that process CAT-ASVAB scores.

It is well known that a below-average examinee looks worse on a reliable test than on an unreliable test. The same is true of entire groups. Suppose a particular subgroup has a mean score below that of the general population. If the test becomes more reliable, the mean score of this subgroup will be even lower. Such adverse impact has nothing to do with "bias" in the usual sense; it is a direct consequence of the increase in the reliability of the test.

METHODOLOGY

Five ASVAB subtests were simulated - General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), and Mathematics Knowledge (MK). The simulation used the three-parameter logistic model of item response theory. Estimated item parameters for two CAT-ASVAB and two PP ASVAB forms were taken from those provided by the Navy Personnel Research and Development Center. Correlations among abilities were based on those in the 1980 reference population. Mean abilities for the minority subgroup were taken to be one standard deviation below those of the general population.

Reliabilities of CAT and PP subtests were given by correlations between raw scores on equivalent forms. CAT subtest scores were converted by equipercentile equating to ASVAB Form 9a raw scores and then to standard scores. A composite score was defined as simply the sum of standard scores on the five subtests.

RESULTS AND DISCUSSION

The CAT version was found to have higher reliability than the PP version for all subtests. As a result, the standard deviation of the composite score was 3 percent higher with CAT than with PP. Mean standard scores for the subtests in the minority sample were lower on CAT than on the PP version by less than a tenth of a standard deviation.

While the results were as expected, their magnitudes suggest that such changes will not present problems when CAT-ASVAB is implemented. Assumptions of item response theory are sure to be violated in real data, and this will reduce reliabilities of CAT subtests below those obtained in simulations. Even with simulated data, increased reliabilities had fairly small effects on the variance of the composite score, and on mean scores of the minority group. (Changes in mean standard scores were less than a tenth of the standard deviation in the Reference Population.) Even smaller differences are likely to be observed in real data.

TABLE OF CONTENTS

INTRODUCTION

Item response theory and fast microcomputers have made computerized adaptive testing (CAT) a reality. In conventional paper-pencil (PP) testing, the same items are administered to all examinees. Items that are too easy or too difficult for a particular examinee provide little information about his ability. In contrast, CAT selects each item using available information about the examinee's ability. As a result, a CAT test can be more reliable than a longer PP test.

If replacement of a PP battery of subtests by a CAT version increases reliabilities of the subtests, the battery is improved. However, such improvement may create some practical problems. One such problem concerns the use of composite scores; the other lies in possible adverse impact on minorities. (These problems are to be expected any time a test battery is made more reliable, not only through the introduction of CAT.)

The purpose of this paper is to illustrate these potential problems by computer simulation of some subtests in the Armed Services Vocational Aptitude Battery (ASVAB). A recent CNA study found that, in an experimental version of CAT-ASVAB, most CAT subtests were more reliable than their PP counterparts [1]. The Department of Defense may implement another CAT-ASVAB in the near future. Therefore, it will be useful to know what can happen when CAT is introduced. In particular, a simulation can be used to study the adverse impact of CAT on minorities even when there is absolutely no bias in the items.

COMPOSITE SCORES

When one form of the ASVAB is replaced by another, the new form of each subtest is equated to the old one, and then the equated raw scores are transformed into standard scores. As a result, variances of subtest standard scores in the two forms are equal.

Selection and classification decisions using the ASVAB are based on composites of subtests rather than single subtests. Therefore, what matters is the equating of composite scores. A new PP form is constructed to be as similar as possible to the old one. As a result, both forms of a subtest are almost equally reliable. Hence, correlations among subtests do not change much from one form to another, and neither do variances of composite scores.

The situation is different with CAT. For each subtest, standard scores based on CAT have the same variance as those based on the reference PP Form 8a. However, if CAT subtests are more reliable, they have higher intercorrelations. This results in CAT composites having larger variances than PP composites in spite of making variances equal at the subtest level. Unequal variances may make it necessary to carry out another equating, now at the composite level. If such an additional

equating does become necessary, it will be a major departure from past practice, requiring software changes on the part of any agency that has to compute composite scores from CAT data. A simulation can provide an idea of what the change in composite variance from PP to CAT may be.

ADVERSE IMPACT

It is well known that a below-average examinee looks worse on a reliable test than on an unreliable test. To make this issue more precise, consider a minority group whose mean standard score on a PP subtest is lower than that of the population as a whole. Suppose the CAT subtest is more reliable, and is equated to the PP subtest using the equipercentile procedure (or a variant of it). Then the mean standard score of the minority population will be lower on the CAT subtest than on the PP version. This result is illustrated in appendix A with a simple numerical example.

It is important to note that such adverse impact has nothing to do with "bias" in the items, in the scoring procedure, or in anything else. As illustrated later in this paper, such impact occurs even in simulated data where the items have the same characteristics in the two groups. It is purely a result of making the subtest more reliable.

METHODOLOGY

Five ASVAB subtests were simulated: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), and Mathematics Knowledge (MK). Distribution of ability in the general population was taken to be standard normal for each subtest. Correlations among the subtests in the 1980 reference population were taken from Maier and Sims [2], corrected for attenuation, and used as correlations among true abilities. Item pools for two CAT forms were those to be used in the Accelerated CAT-ASVAB Project (ACAP); so were subtest lengths: 10 items for PC and 15 for all others. PP ASVAB forms simulated were 9a and 10a. Parameter estimates for all of these items, based on the three-parameter logistic model, were provided by the Navy Personnel Research and Development Center.

A standard normal prior distribution of ability and Owen's approximation for the Bayesian ability estimate [3] were used to score the item responses. Each item was selected for maximum information, subject to exposure control as in ACAP [4]. The Owen estimate at the end of a subtest was converted into a number-correct score on PP Form 9a, linearly transformed, and rounded so that the "raw" CAT score for later analyses was an integer between 0 and 99. The 0-99 scale and Owen's approximation as the final ability estimate were chosen to simplify the computations. Divgi [5] has found that the Owen estimate is as reliable as the correct Bayesian posterior mode and mean.

Each of 2,500 simulated examinees from the general population was administered CAT and both PP forms of the ASVAB. Reliabilities were given by correlations between raw scores on equivalent forms.

Score on CAT Form 1 were transformed into standard scores via equipercentile equating to PP Form 9a. A composite score for each person from the general population was computed as the sum of standard scores on the five subtests. (This does not correspond to any operational ASVAB composite; it is used only as a simple illustration.) The standard deviation of this sum was computed separately for the PP and CAT versions.

The ability distribution in the minority population was similar to that in the general population, except that mean ability on each subtest was -1 (i.e., one standard deviation below the general mean). The sample size again was 2,500. Only CAT Form 1 and PP Form 9a were simulated for each minority examinee. CAT scores were converted to standard scores using the equating already performed. Mean subtest standard scores were computed for both versions.

RESULTS

The resulting reliabilities of PP and CAT subtests in the general sample are presented in table 1. CAT was more reliable for all subtests. The standard deviation of the sum of standard scores was 38.78 for PP and 39.93 for CAT. Thus, the increases in reliabilities from PP to CAT led to a 3-percent increase in the standard deviation of the composite.

TABLE 1

RELIABILITIES OF SUBTESTS

	Subtest				
Version	GS	AR	WK	PC	MK
CAT	.897	.920	.921	.850	.938
PP	.823	.883	.908	.765	.843

Mean CAT and PP standard scores in the minority sample are presented in table 2. They show that increased reliabilities lead to lower mean scores for minority applicants, as expected. The differences are smaller than a tenth of a standard deviation of standard scores in the 1980 Reference Population, which is 10.

TABLE 2

MEAN STANDARD SCORES OF MINORITY EXAMINEES

	Subtest				
Version	GS	AR	WK	PC	MK
CAT	40.45	43.04	41.11	39.61	42.25
PP	41.22	43.39	41.50	40.53	43.14

CONCLUSIONS

The results of this study show that, if all assumptions of the three-parameter logistic model hold, the five subtests are more reliable in the CAT version than in the PP form. In practice, because the model cannot be strictly valid, CAT reliabilities are almost certain to be smaller. Therefore, so will be their impact on the standard deviations of composites and on mean scores of minority applicants. It seems likely, however, that any such effects observed in real data will be small enough to be ignored.

It is important to note that the adverse impact seen in table 2 has nothing to do with "bias." The impact is purely a result of increased reliability. It is an entirely different question whether CAT items have the same characteristics in different subpopulations; that question can only be addressed by analyzing real data.

REFERENCES

[1] CNA Research Contribution 581, *Estimating Reliabilities of CAT-ASVAB Subtests*, by D. R. Divgi, Jan 1988

[2] CNA Report 116, *The ASVAB Score Scales: 1980 and World War II*, by Milton H. Maier and William H. Sims, Jul 1986

[3] Roger J. Owen. "A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing." *Journal of the American Statistical Association* (Jun 1975): 351-356

[4] J. B. Sympson and R. D. Hetter, *Controlling Item-Exposure Rates in Computerized Adaptive Testing*, a paper presented at the annual meeting of the Military Testing Association, Oct 1985

[5] CNA Research Memorandum 87-161, *Properties of Some Bayesian Scoring Procedures for Computerized Adaptive Tests*, by D. R. Divgi, Aug 1987

APPENDIX A

ADVERSE IMPACT OF A MORE RELIABLE TEST

APPENDIX A

ADVERSE IMPACT OF A MORE RELIABLE TEST

Suppose a short old form of a subtest is to be replaced by a new form with twice the number of items and with the items similar to those in the old form. Both are administered to the 1980 reference population. Raw (i.e., number-correct) scores on each form are transformed linearly so that the standard scores have mean 50 and standard deviation 10. By doing so, linear equating of the two forms has been carried out implicitly.

Suppose the reliability of the old form, i.e., the ratio of variances of true and observed scores, is .64. Then its true standard score has a standard deviation of 8 points. The Spearman Brown formula yields the reliability of the new form as .78 [A-1], so its true scores have standard deviation of 8.8.

For any given person, the true score is the same on both forms if it is expressed as the proportion of items answered correctly. (This is the case because the two forms measure exactly the same trait.) Hence the standardized true score (i.e., true score minus population mean divided by standard deviation) is the same on both forms no matter which score scale is used--proportion-correct, number-correct, or standard score--because transformations between these scales are linear. As this is true of every individual, it is also true of the mean for an entire minority subgroup.

Suppose a minority subgroup has mean true score of 42 on the old form. This is below the mean of the general population by one standard deviation (of true scores on the old form). On the new form, the true score one standard deviation below the general mean is 41.2. Thus, due to the greater reliability of the new form, which increases the standard deviation of true scores, the minority mean will be lower by .8 standard score point than with the old form.

REFERENCE

[A-1] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison Wesley, 1968

27 880171.00